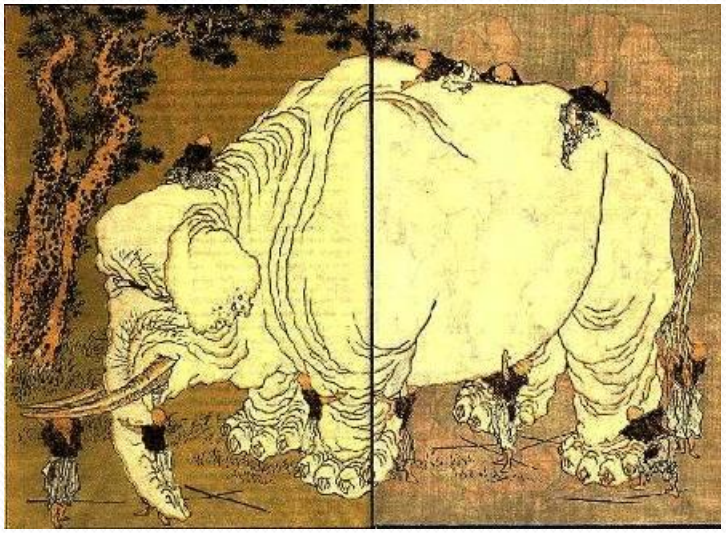
Berkeley Water Center Eco-Science Infrastructure

Deb Agarwal, Catharine van Ingen (Microsoft), Bora Beran (Microsoft), Monte Goode, Keith Jackson, and Robin Weber

Ecological Data Avalanche/Landslide/Tsunami

The era of remote sensing, cheap ground-based sensors and web service access to agency repositories is here

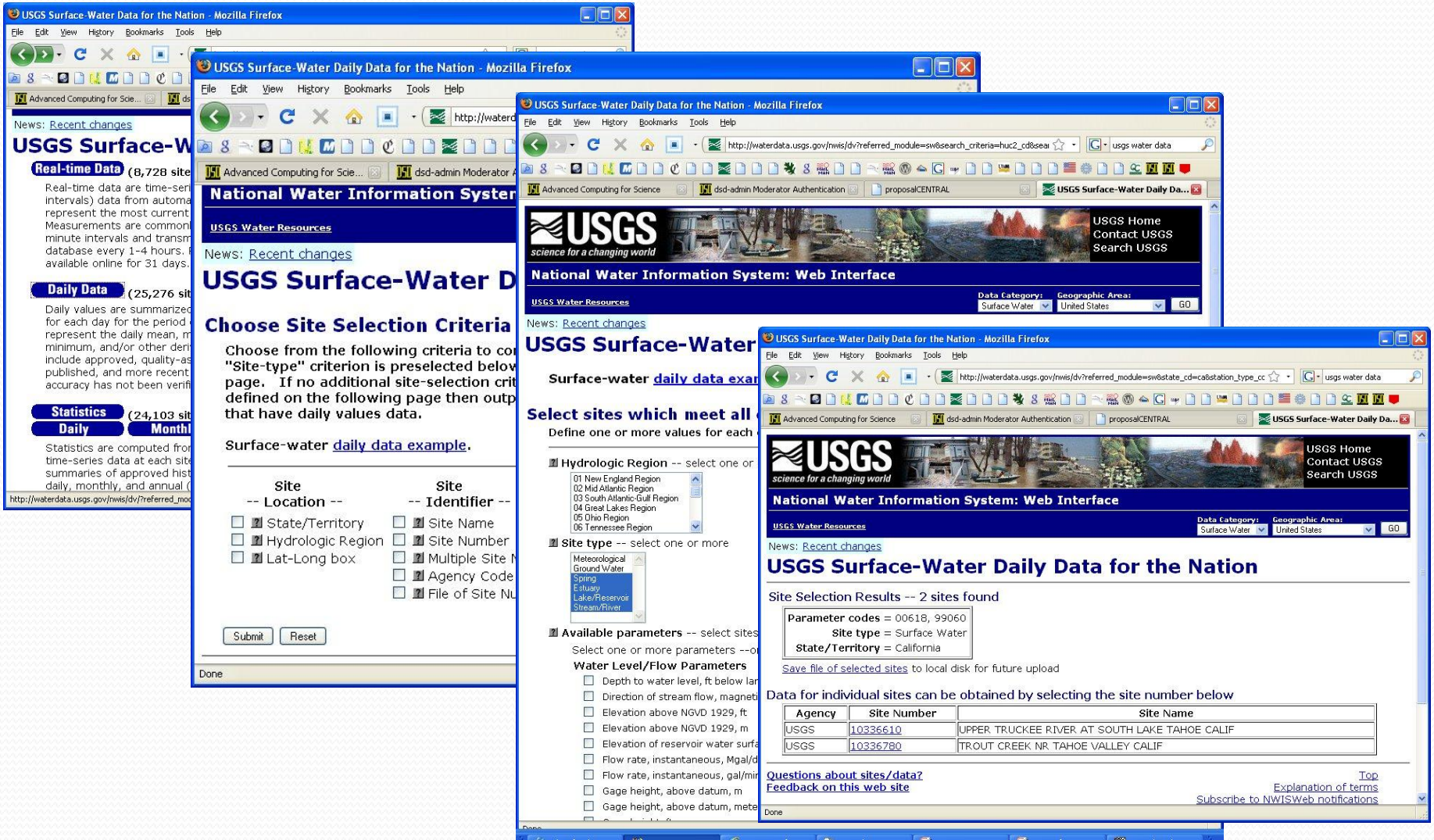
- Extracting and deriving the data needed for the science remains problematic
  - Specialized knowledge
  - Finding the right needle in the haystack



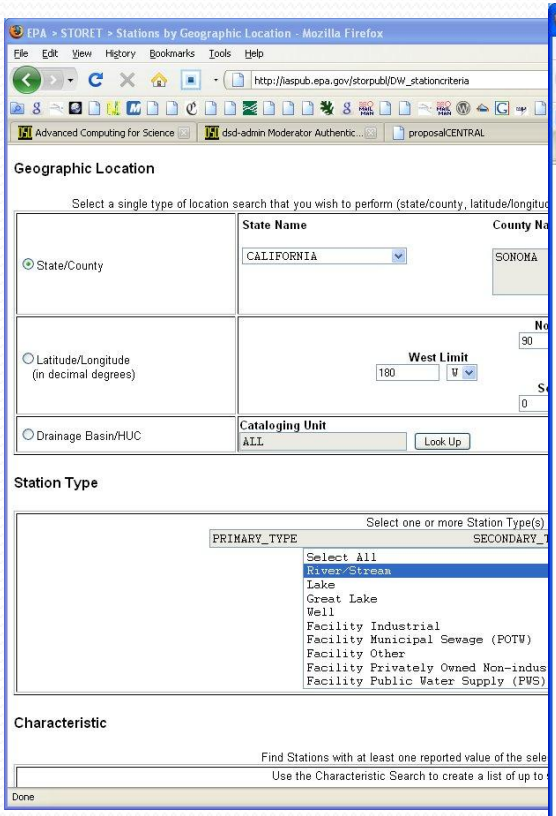
Data Availability is Often no Longer the Problem

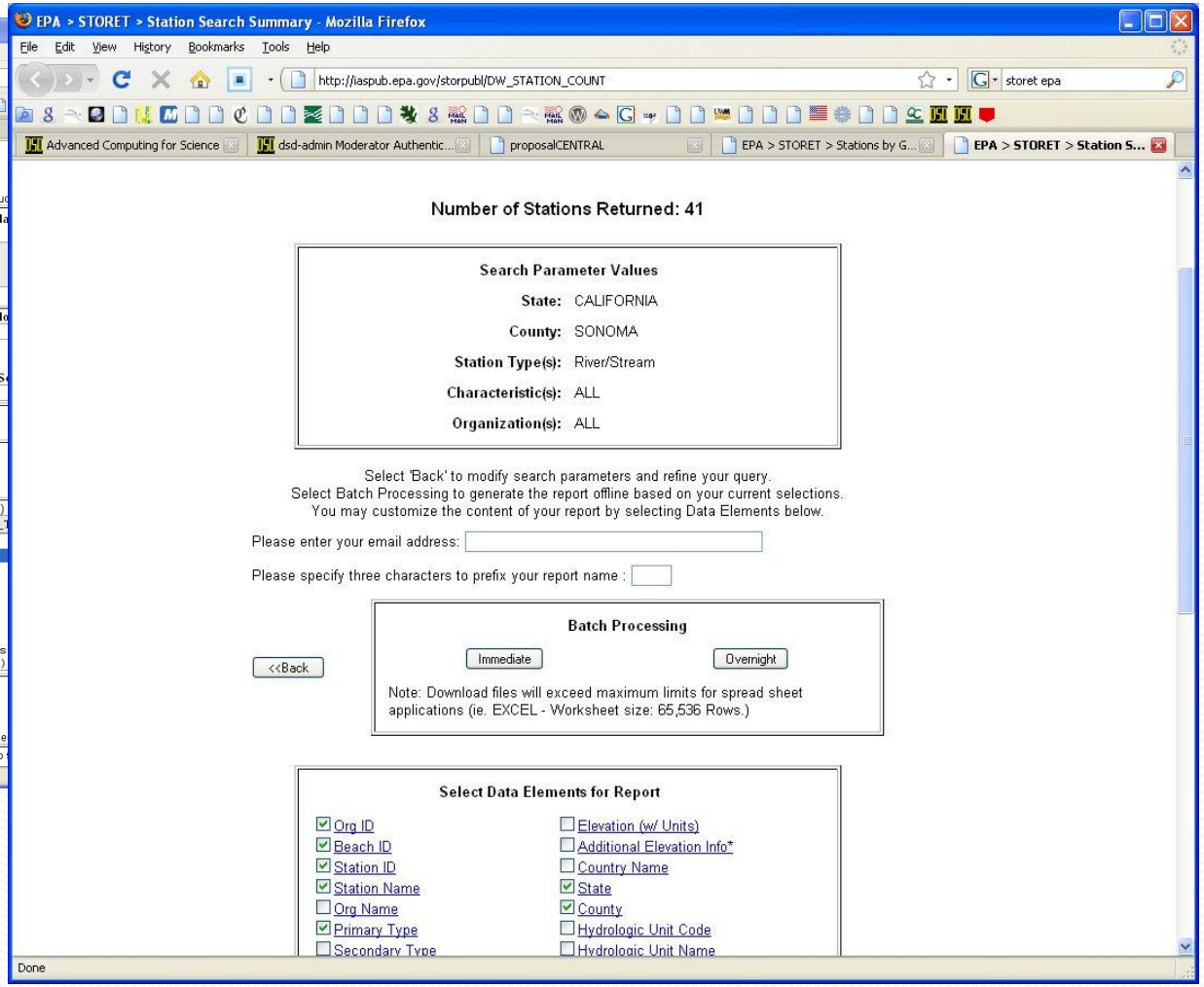
- Broad range of agencies make data available via the web
  - Wide diversity of interfaces
  - Different data formats
  - Different representations
  - Interfaces difficult to navigate
  - Different periods of record
- Gaps in data and diverse periods of record
- Research groups typically use flat files on an internal server
- Data continually updated

USGS NWIS Data Interface



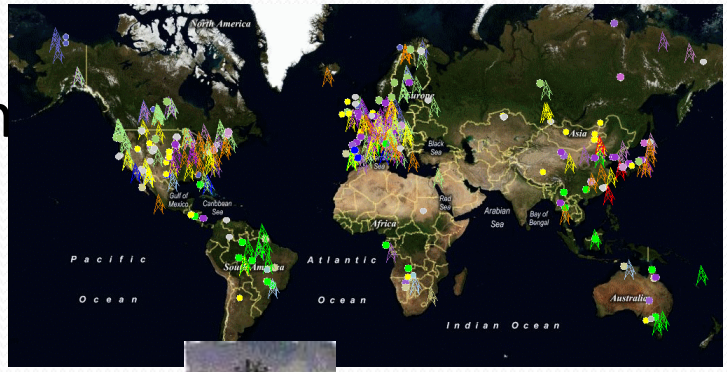
EPA STORET


EPA > STORET > Stations by Geographic Location - Mozilla Firefox
Geographic Location
State/County
Latitude/Longitude (in decimal degrees)
Drainage Basin/HUC
Station Type
Characteristic

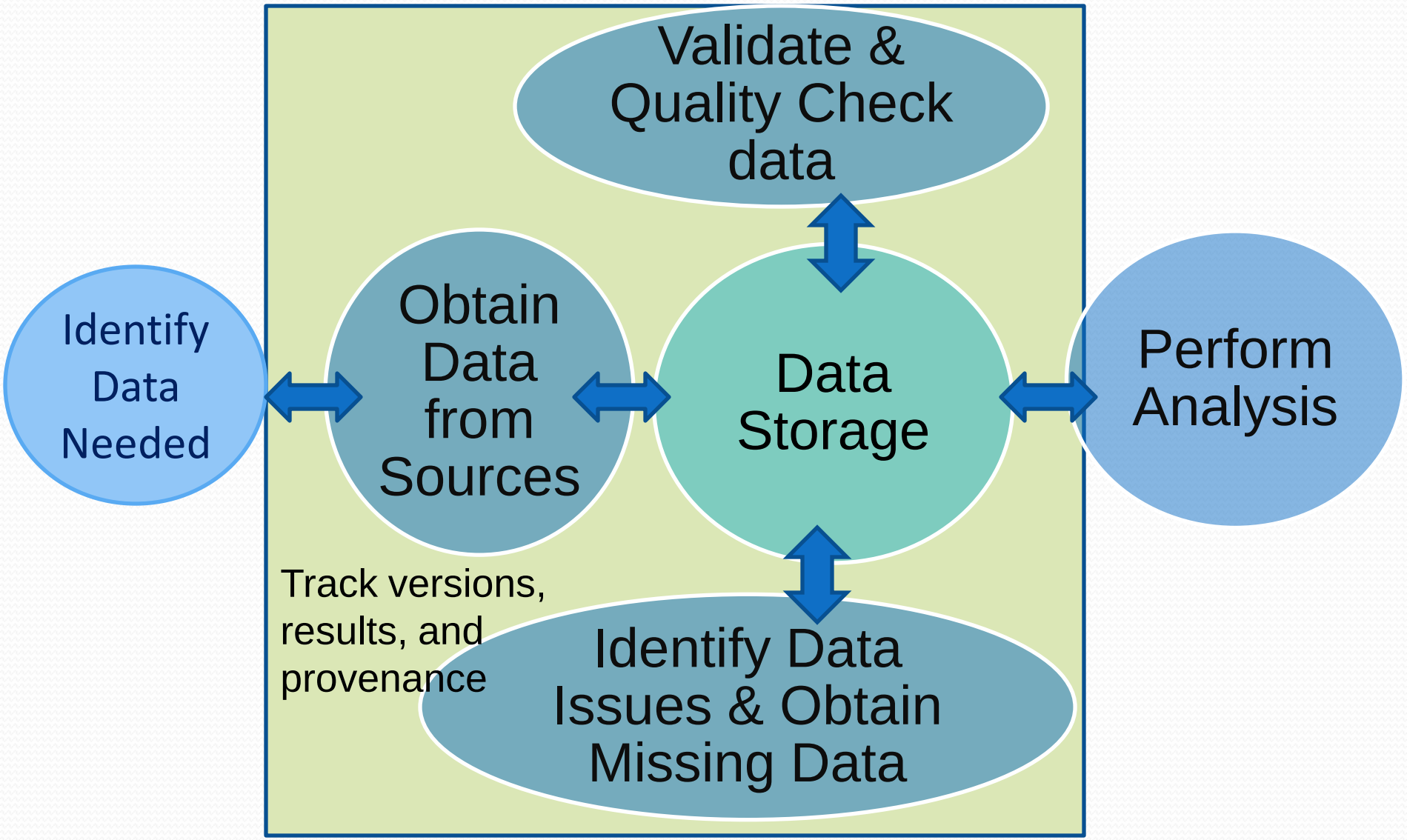

EPA > STORET > Station Search Summary - Mozilla Firefox
Number of Stations Returned: 41
Search Parameter Values
State: CALIFORNIA
County: SONOMA
Station Type(s): River/Stream
Characteristic(s): ALL
Organization(s): ALL
Select 'Back' to modify search parameters and refine your query.
Select Batch Processing to generate the report offline based on your current selections.
You may customize the content of your report by selecting Data Elements below.
Please enter your email address:
Please specify three characters to prefix your report name :
<<Back
Batch Processing
Immediate
Overnight
Note: Download files will exceed maximum limits for spread sheet applications (ie. EXCEL - Worksheet size: 65,536 Rows.)
Select Data Elements for Report
Org ID
Beach ID
Station ID
Station Name
Org Name
Primary Type
Secondary Type
Elevation (w/ Units)
Additional Elevation Info*
Country Name
State
County
Hydrologic Unit Code
Hydrologic Unit Name

Example Experiment Dataset - FLUXNET

- Global carbon flux and micrometeorological measuremen
  - 960 site years of data
    - 40 variables and 60 quality tags
    - 1/2 hourly measurements
  - 250 sites
  - 160 site investigators (collecting data)
  - 65 synthesis paper proposals involving ~130 collaborators
- Goal: Understand the role of photosynthesis in carbon cycle

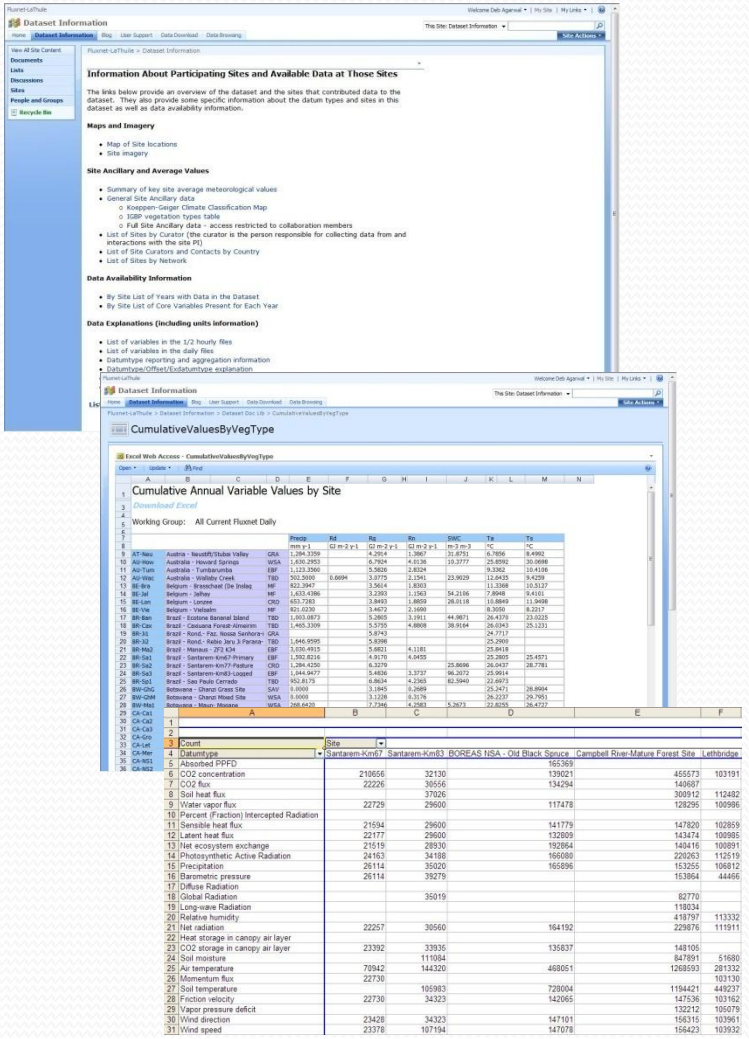


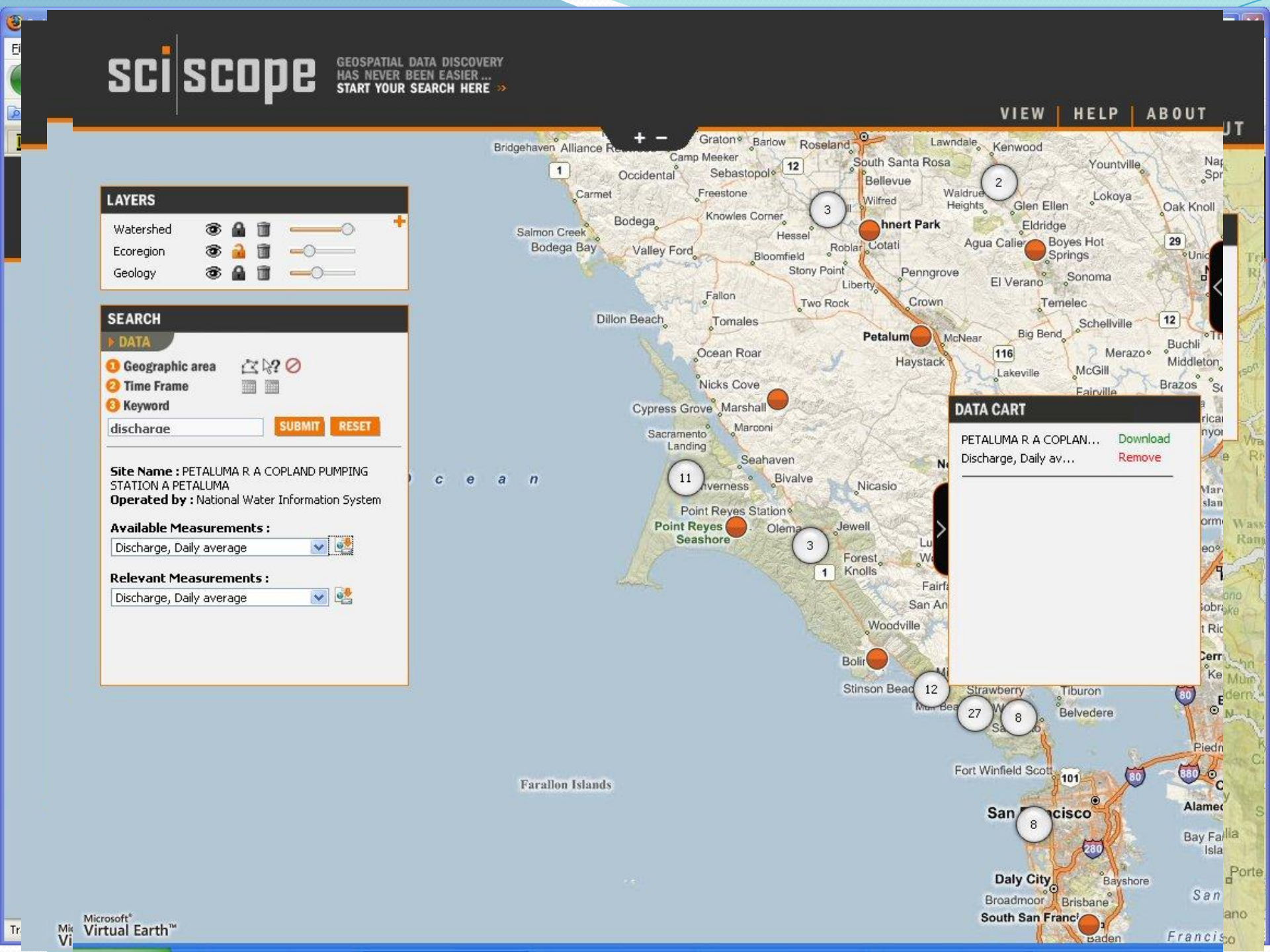
Target Data Analysis Flow



Developing Common Eco-Science Data Infrastructure

- Graphical data selection tool
- Data ingested into a database
- Database schema organized by dataset, site, time, and variable
- On-line analytical processing (OLAP) data cubes provide browsing access
- Reports provide web pages describing the data
- Collaboration tools facilitate communication between data providers and data users as well as between data users



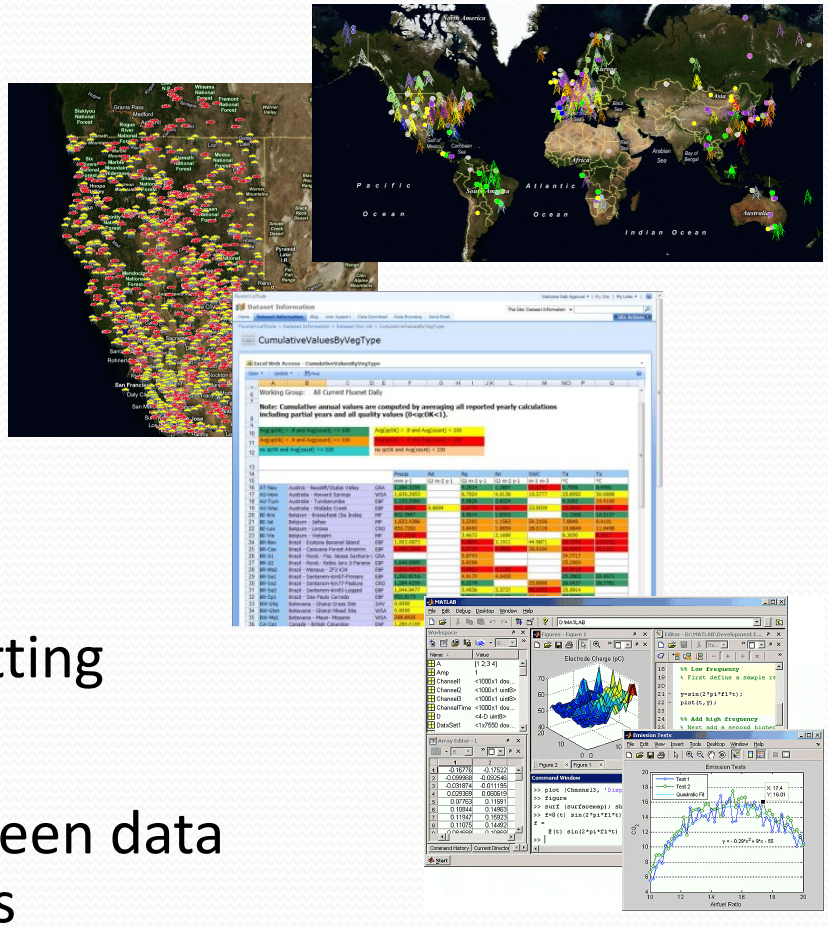

sciscope
GEOSPATIAL DATA DISCOVERY
HAS NEVER BEEN EASIER ...
START YOUR SEARCH HERE »
VIEW | HELP | ABOUT
LAYERS
Watershed
Ecoregion
Geology
SEARCH
DATA
1 Geographic area
2 Time Frame
3 Keyword
discharge
SUBMIT
RESET
Site Name : PETALUMA R A COPLAND PUMPING STATION A PETALUMA
Operated by : National Water Information System
Available Measurements :
Discharge, Daily average
Relevant Measurements :
Discharge, Daily average
DATA CART
PETALUMA R A COPLAN...
Download
Discharge, Daily av...
Remove
Petaluma
Point Reyes Seashore
Farallon Islands
Microsoft Virtual Earth™

Behind SciScope

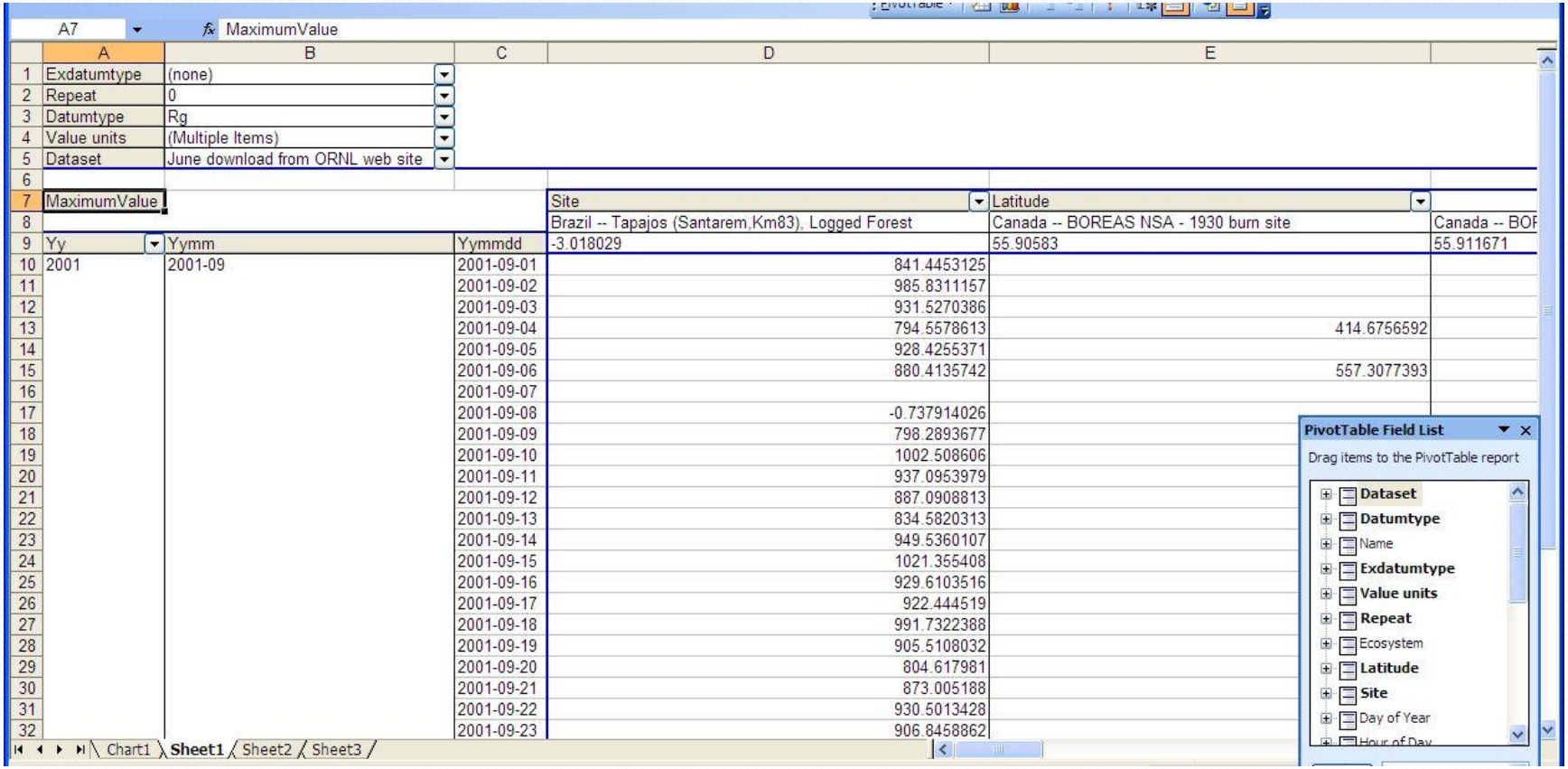
- Contains information on where, when (time frame) and what is being measured for about 1.65 million sites in the US
- Scraped/crawled on a regular basis
- SciScope currently hosts only metadata.
- Data are requested on the fly from the original publisher using web service wrappers written specifically for each data source.
- Data are reformatted to provide a unified view over the repositories.

Interaction and Visualization of Data

- Mashup maps
  - locate sensors
- Data summary reports
  - Data browsing
  - Data selection
- OLAP data cube interfaces
  - Matlab – analysis and plotting
  - Excel – data browsing and plotting
- Collaboration capabilities
  - Improve communication between data producers and data consumers



Pivot Tables



A7 MaximumValue

	A	B	C	D	E
1	Exdatumtype	(none)			
2	Repeat	0			
3	Datumtype	Rg			
4	Value units	(Multiple Items)			
5	Dataset	June download from ORNL web site			
6					
7	MaximumValue			Site	Latitude
8				Brazil -- Tapajos (Santarem,Km83), Logged Forest	Canada -- BOREAS NSA - 1930 burn site
9	Yy	Yymm	Yymmdd	-3.018029	55.90583
10	2001	2001-09	2001-09-01	841.4453125	
11			2001-09-02	985.8311157	
12			2001-09-03	931.5270386	
13			2001-09-04	794.5578613	414.6756592
14			2001-09-05	928.4255371	
15			2001-09-06	880.4135742	557.3077393
16			2001-09-07		
17			2001-09-08	-0.737914026	
18			2001-09-09	798.2893677	
19			2001-09-10	1002.508606	
20			2001-09-11	937.0953979	
21			2001-09-12	887.0908813	
22			2001-09-13	834.5820313	
23			2001-09-14	949.5360107	
24			2001-09-15	1021.355408	
25			2001-09-16	929.6103516	
26			2001-09-17	922.444519	
27			2001-09-18	991.7322388	
28			2001-09-19	905.5108032	
29			2001-09-20	804.617981	
30			2001-09-21	873.005188	
31			2001-09-22	930.5013428	
32			2001-09-23	906.8458862	

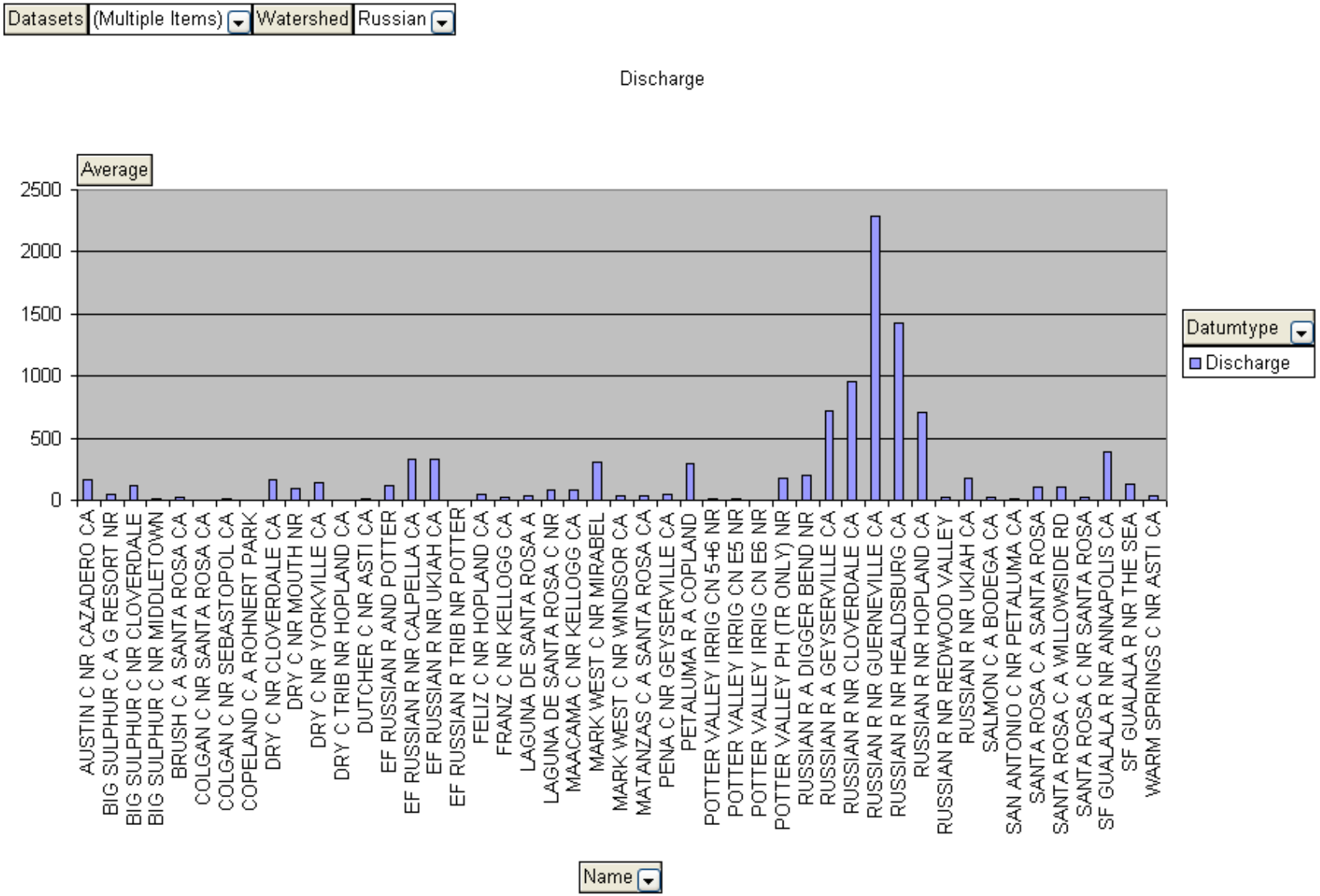
Chart1 | Sheet1 | Sheet2 | Sheet3

PivotTable Field List

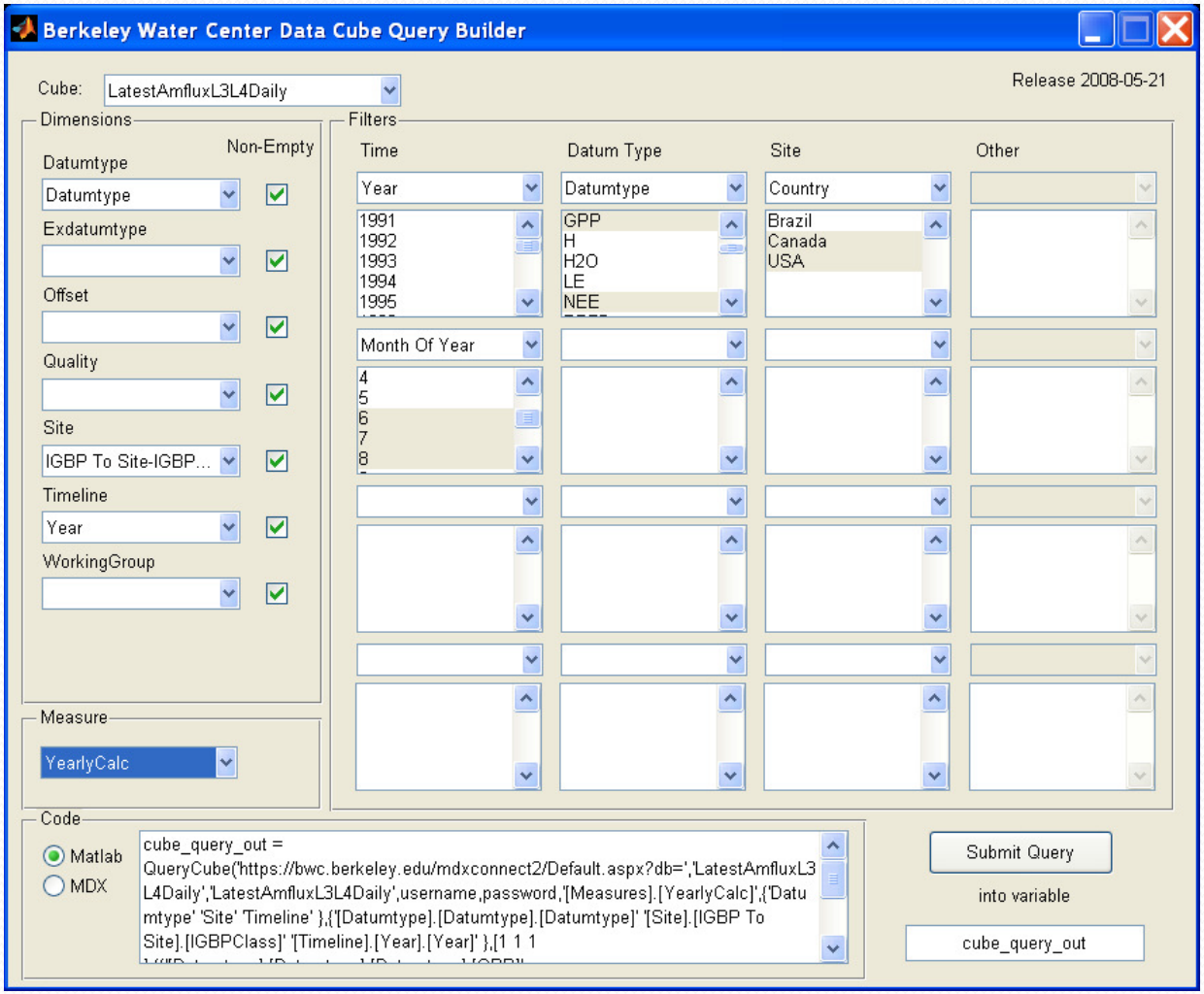
Drag items to the PivotTable report

- Dataset
- Datumtype
- Name
- Exdatumtype
- Value units
- Repeat
- Ecosystem
- Latitude
- Site
- Day of Year
- Hour of Day

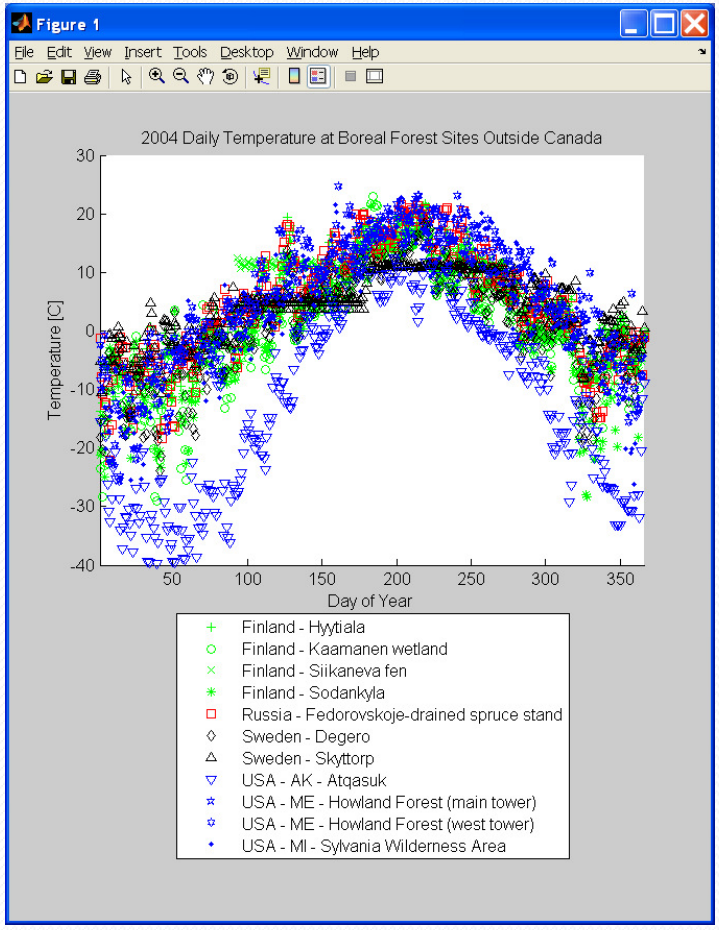
Pivot Charts



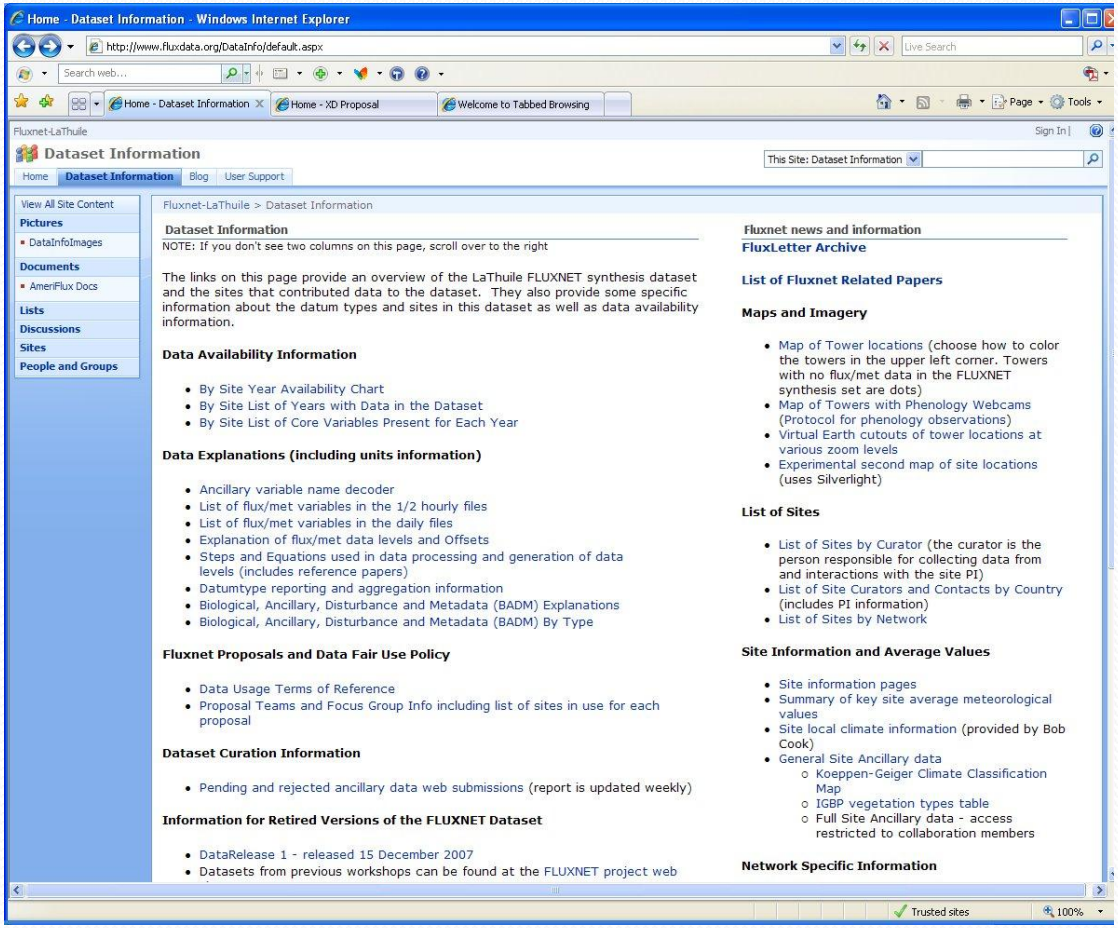
MatLab Interface



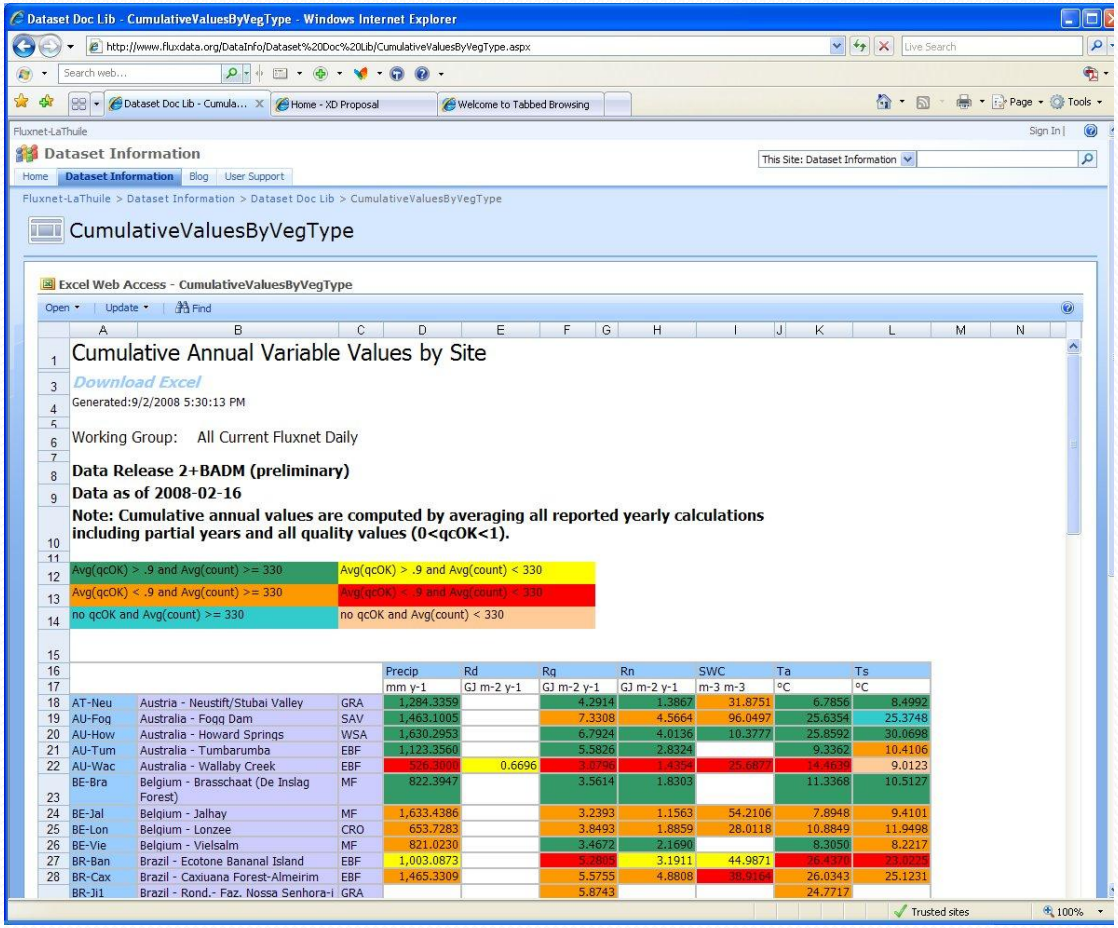
Plotting Results



SharePoint Portal



SharePoint Hosted Reports



Acknowledgements

Berkeley Water Center, University of California, Berkeley, Lawrence Berkeley Laboratory
- Jim Hunt
- Deb Agarwal
- Robin Weber
- Monte Good
- Rebecca Leonardson (student)
- Matt Rodriguez
- Carolyn Remick
- Susan Hubbard

University of Virginia
- Marty Humphrey
- Norm Beekwilder

Microsoft
- Catharine van Ingen
- Bora Beran
- Jayant Gupchup (student)
- Savas Parastidis
- Andy Sterland
- Nolan Li (student)
- Tony Hey
- Dan Fay
- Jing De Jong-Chen
- Stuart Ozer
- SQL product team
- Jim Gray

Ameriflux Collaboration
- Beverly Law
- Youngryel Ryu (postdoc)
- Tara Stiefl (student)
- Gretchen Miller (student)
- Mattias Falk
- Tom Boden
- Bruce Wilson

Fluxnet Collaboration
- Dennis Baldocchi
- Rodrigo Vargas (postdoc)
- Dario Papale
- Markus Reichstein
- Bob Cook
- Susan Holladay
- Dorothea Frank

National Marine Fisheries Service
Sonoma Ecology Center

The saga continues at
http://bwc.berkeley.edu and
http://www.fluxdata.org